The Journal of Experimental Education
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/vjxe20

Five Methods to Score the Teacher Observation of Classroom Adaptation Checklist and to Examine Group Differences

Ze Wang[a], David Rohrer[a], Chi-ching Chuang[a], Mayo Fujiki[a], Keith Herman[a] & Wendy Reinke[a]
[a] University of Missouri
Published online: 28 Mar 2014.

To cite this article: Ze Wang, David Rohrer, Chi-ching Chuang, Mayo Fujiki, Keith Herman & Wendy Reinke (2015) Five Methods to Score the Teacher Observation of Classroom Adaptation Checklist and to Examine Group Differences, The Journal of Experimental Education, 83:1, 24-50, DOI: 10.1080/00220973.2013.876230

To link to this article: http://dx.doi.org/10.1080/00220973.2013.876230

THE JOURNAL OF EXPERIMENTAL EDUCATION, *83*(1), 24–50, 2015

ISSN: 0022-0973 print/1940-0683 online
DOI: 10.1080/00220973.2013.876230

Routledge
Taylor & Francis Group

Five Methods to Score the Teacher Observation of Classroom Adaptation Checklist and to Examine Group Differences

Ze Wang, David Rohrer, Chi-ching Chuang, Mayo Fujiki, Keith Herman, and Wendy Reinke
University of Missouri

This study compared 5 scoring methods in terms of their statistical assumptions. They were then used to score the Teacher Observation of Classroom Adaptation Checklist, a measure consisting of 3 subscales and 21 Likert-type items. The 5 methods used were (a) sum/average scores of items, (b) latent factor scores with continuous indicators, (c) latent factor scores with ordered categorical indicators using the mean- and variance-adjusted weighted least squares estimation method, (d) latent factor scores with ordered categorical indicators using the full information maximum likelihood estimation method, and (e) multidimensional graded response model using the Bock-Aitkin expectation-maximization estimation procedure. Measurement invariance between gender groups and between free/reduced-price lunch status groups was evaluated with the second, third, fourth, and fifth methods. Group mean differences based on the 5 methods were calculated and compared.

Keywords *factor analysis, measurement invariance, ordered categorical variables, scoring methods, TOCA-C*

IN MANY EDUCATIONAL and behavioral studies, participants' responses and behaviors are recorded and then scored. The scoring procedure may follow some predefined rubrics (e.g., Woodcock-Johnson III has detailed scoring procedures, McGrew & Woodcock, 2001), or be carried out along with statistical analysis that involve explicit or implicit assumptions (e.g., Cooke, Kosson, & Michie, 2001, examined latent means using two different techniques). Different scoring methods may result in different conclusions about the same research hypothesis. Therefore, the scoring method used in a particular study is an important psychometric issue. In some cases, the scoring method may be obvious and easy to understand such as when the total number of correct answers on a test is recorded. In other cases, the scoring method may be complicated such as when people to be compared respond to different test items (e.g., the National Assessment of Educational Progress uses a matrix sampling technique where students responded to different item blocks, see Mislevy, Johnson, & Muraki, 1992).

Address correspondence to Ze Wang, Department of Educational, School and Counseling Psychology, University of Missouri, 16 Hill Hall, Columbia, MO 65211, USA. E-mail: wangze@missouri.edu

When responses are categorical, from a canonical analysis perspective, McKeon (1966) pointed out that scoring methods involve considerations of the context/nature of variables (e.g., each person responding to all items, or each person rating responses), response/form of data (e.g., Likert-type where a single response is recorded for any item by a person, or Thurstone data when a person selects statements from a list), and quantification method (e.g., how to assign weights to different responses). Lord and Novick (1968) stated:

> . . . measurement (or scaling) is a fundamental part of the process of theory construction. A major problem of mental test theory is to determine a good interval scaling to impose when the supporting psychological theory implies only ordinal properties (p. 22).

Efforts to "determine a good interval scaling" (i.e., how to score responses) are probably best shown by researchers of item response theory (IRT; e.g., Hambleton & Swaminathan, 1985). Given the more recent integration of IRT with other modeling techniques, methodological developments in techniques to model (ordered) categorical data are in line with these efforts. For example, item factor analysis can be situated in the confirmatory factor analysis (CFA) framework and the IRT framework (Wirth & Edwards, 2007) .[1] In drastic contrast to methodological development is the continuing practice of using scoring methods that are inconsistent with the supporting psychological theory and/or the mathematical system that the measurement level is associated with (see Byrne, 2010, for a discussion on practices in testing latent mean differences, p. 231).

In this study, we compared five different scoring methods in terms of their statistical assumptions. The five methods use (a) sum/average scores of items, (b) latent factor scores with continuous indicators, (c) latent factor scores with ordered categorical indicators using the mean- and variance-adjusted weighted least squares (WLSMV) estimation method, (d) latent factor scores with ordered categorical indicators using the full information maximum likelihood (FIML) estimation method, and (e) multidimensional graded response model using the Bock-Aitkin expectation-maximization (BAEM) estimation procedure. With an empirical example, we examined how the scoring methods resulted in conclusions in terms of statistically significant group differences and in terms of the magnitude (i.e., effect sizes) of those differences. It is also worth noting that our study does not intend to investigate systematically which method is a better one under certain conditions; such investigation requires simulation studies that are not the focus of this study.

FIVE SCORING METHODS

Sum/Average Scores of Items

One of the most common methods to score individuals in psychological and behavioral domains is to sum/average raw scores corresponding to items that load on a latent factor based on factor analysis (e.g., Harter, 1981). Usually, factor analysis has been done by the developer of the instrument and items are clustered. Later, users of the instrument calculate the sum or average of raw scores. Sometimes, reverse coding of items is necessary before the summation/averaging.

[1]Not every IRT model has a CFA counterpart for item factor analysis (see Bollen, Bauer, Christ, & Edwards, 2010).

When the sum/average scores are used to examine differences in the psychological construct that those scores supposedly represent, a distributional assumption (e.g., normal distribution) is usually made on the sum/average scores to facilitate statistical inference and testing this distributional assumption is not difficult. However, implicit assumptions about *items* are usually ignored. Reporting the reliability value that is calculated with item raw scores from the same data—instead of the value reported by earlier instrument developers—is helpful. For example, Cronbach's alpha assumes true-score equivalent items with uncorrelated residuals (i.e., essentially tau-equivalent items; McDonald, 1999; however, see Raykov, 2012, for a discussion about Cronbach's alpha and reliability). True-score equivalent items require that item raw scores have the same true score but item residual variances may differ. From a CFA perspective, this means that all items have the same factor loadings. However, it is important to point out that true-score equivalent items do not require a normality assumption on the distribution of residual variances, which is a usual assumption in CFA.

Latent Factor Scores with Continuous Indicators

When the items are not true-score equivalent, the true score part of the composite score as the sum/average of item scores receives different contributions from different items. Therefore, items differ in their contributions to the examination of individual differences in the psychological construct. Such item differences reflect a shortening or lengthening factor of the measurement scale from one item to another. In factor analysis, differences in factor loadings for items may indirectly indicate different contributions of items (here, we use *indirectly* because factor loadings themselves are not weights assigned to items to form a composite score). The basic equation can be written as follows:

$$\mathbf{Y}_i = \boldsymbol{\alpha} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i^T + \boldsymbol{\varepsilon}_i \tag{1}$$

where

$\mathbf{Y}_i$ is a vector representing person i's responses to all J items, $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{iJ})$.

$\boldsymbol{\alpha}$ is a vector of intercepts for all J items, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_J)$.

$\boldsymbol{\Lambda}$ is a J-by-M matrix of factor loadings where M is the number of latent factors. When each item loads only on one latent factor, there is only one nonzero element in each row of the matrix $\boldsymbol{\Lambda}$.

$\boldsymbol{\eta}_i$ is a vector of dimension M, representing latent factor scores for person i.

$\boldsymbol{\varepsilon}_i$ is a vector of measurement error of all J items for person i.

For model identification, some parameters in Equation (1) have to be fixed. For an easy guide on identification for CFA models, see Brown (2006); for a more technical guide involving multiple groups, see Millsap and Olivera-Aguilar (2012).

Using latent factor scores as the scoring method to examine differences in the psychological construct seems viable (e.g., Steinmetz, Schmidt, Tina-Booh, Wieczorek, & Schwartz, 2009, used this method to examine latent mean differences between educational groups). With this scoring method, item scores are assumed on an interval scale (i.e., continuous indicators). Because item scores usually do not meet the multivariate normality assumption required by the commonly used

maximum likelihood estimation, some corrections are necessary. One correction that has shown to be robust is the Yuan-Bentler adjusted test statistic (Yuan & Bentler, 2000). The maximum likelihood estimator with robust standard errors and adjusted chi-square statistics (MLR) in Mplus (L. K. Muthén & B. O. Muthén, 2012) is asymptotically equivalent to the Yuan-Bentler adjusted test statistic and has the additional advantage of taking into consideration nonindependence of observations with complex data structures.

Latent Factor Scores With Ordered Categorical Indicators Using WLSMV Estimation

The aforementioned scoring method assumes that items are on an interval scale and ignores the ordered categorical nature of Likert-type items. An interval scale "assigns meaning not only to scale values and their relative order but also to relative differences of scale values. In effect, an interval scale specification establishes a distance function over all pairs of elements . . . " (Lord & Novick, 1968, p. 21). When Likert-type items are used, this means equal distance between response options when equidistance real numbers are assigned to those response options. When such an equal distance between responses assumption is violated, different estimation methods that are appropriate for ordinal variables can be used. Earlier research of the weighted least squares estimator (WLS), whose continuous variable counterpart is the asymptotically distribution free estimator, takes a latent response variable approach where thresholds are used to form ordered categories corresponding to observed responses (Jöreskog, 1990; Muthén, 1984). Later, more robust WLS estimators take a similar approach and incorporate different weight matrixes. Because those estimators take into consideration the ordered nature, they seem appropriate for analysis with Likert-type items. Latent factor scores obtained from such analysis provides another scoring method (for example, Dorman, 2003, applied this method to examine differences in the What Is Happening in This Class? questionnaire across countries, grade levels, and student gender). The WLS and robust WLS estimators are limited information estimation methods in that the factor model is tested against a correlation matrix for latent response variables (Edwards, Wirth, Houts, & Xi, 2012). Those correlation matrixes (tetrachoric for dichotomous items and polychoric for ordered-categorical items with more than two categories) are obtained with the assumption that latent response variables are bivariate normal. In Mplus, those correlations are estimated using a two-stage procedure described by Olsson (1979; a one-stage procedure is also described). Muthén (1984) used the method of collapsing categories for variables to avoid low bivariate frequencies in Pearson chi-square tests of bivariate normality. Of the available WLS estimators, the WLSMV estimator performs well under different conditions and is recommended (Beauducel & Herzberg, 2006; Flora & Curran, 2004). The scoring method using the WLSMV estimator takes into consideration the ordinality of Likert-type items and uses limited information (i.e., bivariate information is summarized in tetrachoric/polychoric correlations). The following is the basic mathematical representation of the model:

$$\mathbf{Y}_i^* = \boldsymbol{\alpha} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i^T + \boldsymbol{\varepsilon}_i \tag{2}$$

The difference between Equation (2) and Equation (1) is that a vector of latent responses $\mathbf{Y}_i^*$ replaces the vector of observed responses $\mathbf{Y}_i$. The latent responses relate to the observed ones via

a threshold model for item j:

$$Y_{ij} = \begin{cases} 0 \text{ if } -\infty < Y_{ij}^* \leq \tau_{j1} \\ 1 \text{ if } \tau_{j1} < Y_{ij}^* \leq \tau_{j2} \\ \ldots \\ K_j \text{ if } \tau_{jK_j} < Y_{ij}^* < \infty \end{cases}$$

where $\tau_{j1}, \tau_{j2}, \ldots, \tau_{jK_j}$ are threshold parameters for item j.

Latent Factor Scores With Ordered Categorical Indicators Using FIML Estimation

When the factor model is tested against the data (i.e., sample response patterns), instead of summary statistics (e.g., tetrochoric or polychoric correlations), a full-information estimation method can be used (e.g., Edwards, Wirth, Houts, & Xi, 2012, applied this method to items measuring government responsibility). This estimator does not rely on a tetrachoric or polychoric matrix. This approach is equivalent to a 2PL IRT model for dichotomous items or a graded response model for polytomous items using a probit link (Wirth & Edwards, 2007). Factor scores obtained are equivalent to IRT scores (however, a scaling factor D = 1.7 has to be considered if the IRT model is a logistic model rather than a normal ogive model. Their relation is about 1 logit = 1.7 probits). The disadvantage is that it requires numerical integration and is therefore time consuming. Because frequency tables (and response patterns) are used as data information instead of a correlation matrix, Mplus does not provide global model fit statistics such as CFI and RMSEA for this method.

It should be noted that the term *full information* here refers to the fact that each data point (or more accurately, each response pattern) contributes to the estimation procedure, in contrast to the earlier described methods where only summary statistics (e.g., variances and covariances, polychoric correlations) contribute to the estimation. In a different context about missing data on continuous variables, there is also a FIML estimator. In that context, the term *full information* generally refers to the fact that partial data points (cases with missing values on variables) are used in the estimation procedure in addition to complete cases (see Arbuckle, 1996); and that summary statistics (e.g., variances and covariances, polychoric correlations) based on both partial data points and complete cases contribute to the estimation. In this missing data treatment sense, the MLR estimator in the second scoring method is a full information method when missing data are present in the sample. Edwards, Wirth, Houts, and Xi (2012) provided a good description about limited information and full information estimators for categorical data in structural equation modeling. Their differentiation between *limited* and *full* information is consistent with this scoring method used in the present study (i.e., full information means response patterns are modeled and limited information means that summary statistics such as polychoric correlations are modeled).

Ability/Proficiency Scores from Multidimensional IRT Model[2]

Similar to the FIML approach for ordered categorical indicators in CFA, IRT models specify the probability function that a person endorses an item (for dichotomously scored items) or gets credit for his/her response on an item (for polytomously scored items). Undimensional IRT

[2]We thank one reviewer for suggesting this additional scoring method.

models (e.g., 1PL, 2PL, or 3PL) are differentiated by the number of parameters specified for each item. In these models, two sets of parameters are estimated: item parameters and person/group parameters. Item parameters are usually estimated by one of the three maximum likelihood methods: (a) joint maximum likelihood, (b) marginal maximum likelihood, and (c) conditional maximum likelihood. After the item parameters are estimated, an ability/proficiency estimate for each person can be obtained (here, we use ability/proficiency to refer to what the person parameters are estimated for. It could also be attitude, engagement, tendency, and so forth. We are aware that ability/proficiency bears meanings in the assessment of the cognitive domain. IRT models are not limited to the cognitive domain and can be used for psychological and behavioral domains as well). Three popular estimation methods for ability/proficiency parameters are maximum likelihood estimation, maximum a posteriori, and expected a posteriori (Embretson & Reise, 2000; Osterlind & Wang, 2012).

Unidimensional IRT models, in which only one domain (e.g., math proficiency) is modeled, are most often used. Those models allow the examination of individual differences along one dimension. Unidimensionality is considered an essential IRT assumption in those earlier models. More recently, multidimensional IRT models (Reckase, 2009) that include multiple domains are gaining research attention. Of them, within-item multidimensionality models are useful in situations where there is a general domain but items can be grouped in clusters (e.g., testlet models); between-item multidimensionality models are conceptually similar to multifactor CFA models in that each item measures one domain/factor and there are multiple domains/factors. Estimation methods for multidimensional IRT models are available (e.g., Bock & Aitkin, 1981; Cai, 2010; Schilling & Bock, 2005). The major difference between the CFA and IRT frameworks is in the nonlinear link functions used. In the IRT framework, the link function is usually the logit. For CFA models with continuous indicators, the link function is the identity link function. For CFA models with ordered–categorical indicators, the estimation process (e.g., WLSMV) usually consists of multiple stages that involve estimation of the correlation matrixes first (see earlier description). It is conceptually similar to the probit link function since polychoric correlations are usually estimated under the bivariate normality assumption. For the FIML for ordered-categorical indicators in CFA, the probit link is used.

For this multidimensional IRT scoring method in the present study, a between-item multidimensional IRT model (Adams, Wilson, & Wang, 1997), generalized from the graded response model (Samejima, 1969) was used. The mathematical representation of this model is in equation (3).

$$\log\left(\frac{p(Y_{ij}\le k)}{1-p(Y_{ij}\le k)}|\boldsymbol{\theta}_i,\mathbf{a}_j,\mathbf{c}_j\right)=\mathbf{a}_j\theta_i^{\mathrm{T}}+c_{j(k+1)},k=0,1,...,K_j-1 \tag{3}$$

where

Y_{ij} is the response of person i to item j. Response options/ordered categories of item are $0, 1, ..., K_j$.

K_j is the number of thresholds that separate the response options/ordered categories of item j.

$\boldsymbol{\theta}_i$ is the ability vector of dimension M for person i, $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, ..., \theta_{iM})$.

$\mathbf{a}_j$ is the item discrimination vector of dimension M for item j, $\mathbf{a}_j = (a_{j1}, a_{j2}, ..., a_{jM})$. One and only one of $a_{j1}, a_{j2}, ..., a_{jM}$ takes a nonzero value and all the others are equal to zero for a given item j.

$\mathbf{c}_j$ is the intercept vector of dimension K_j for item j, $\mathbf{c}_j = (c_{j1}, c_{j2}, ..., c_{jK_j})$, and $c_{j1} \geq c_{j2} \geq ... \geq c_{jK_j}$.

In this model, each item has one nonzero discrimination parameter and several intercept parameters. The number of intercept parameters for an item is equal to the number of responses minus one (in this study, not all items have the same number of responses in the analysis because of category collapsing). For this model, there is a direct link between the intercept parameters and the threshold parameters that are more commonly used in graded response models: for $a_{jm} \neq 0, m = 1, 2, ..., M$, $c_{j(k+1)} = -a_{jm}b_{j(k+1)}, k = 0, 1, ..., K_j - 1$. The link function is a logit function without the scaling factor $D = 1.7$.

IRTPRO is a new IRT computer program with the capacity for multigroup, multidimensional IRT modeling (Cai, du Toit, & Thissen, 2011). We used IRTPRO 2.1 and the BAEM (Bock & Aitkin, 1981) estimation method in this study.

USING SCORING METHODS TO EXAMINE GROUP MEAN DIFFERENCES

In a sense, the purpose of scoring individuals' responses is to examine differences. A common category of differences in psychological and behavioral domains is group differences in mean scores (here, we discuss only cross-sectional group differences). Group differences observed in a sample are compared with the sampling error under a null hypothesis (the null hypothesis is usually the hypothesis of no difference), and statistical inference can therefore be made. In addition, the magnitude of differences can be quantified and standardized. For group comparisons, because of different statistical assumptions associated with different scoring methods, these methods likely result in differences in statistical inferences and magnitude of group differences. We subsequently describe several important statistical assumptions for group comparison when these scoring methods are used.

With the sum/average scores method, group mean differences are usually examined using independent samples *t* test (for two groups) or analysis of variance (for two or more groups). Besides the normality assumption on the sum/average scores (here, we discuss parametric methods only. There are nonparametric methods that don't require the normality assumption), equal within group variance is preferred when comparing group means and this assumption can be tested. When the equal within group variance is not met, remedies can be made to facilitate group mean comparisons. In addition, it is possible to test for item-level differences (differential item functioning [DIF]) between groups. Statistical techniques for DIF with this method include logistic regression, generalized Mantel-Haenszel, or logistic discriminant function analysis (see Zumbo, 1999, for a description). However, in this study, we focus on group differences at the test level (i.e., group mean differences on the sum/average scores) instead of at the item level. Another commonly used term for test level differences is test bias. There has been extensive research on bias in cognitive ability tests, especially standardized tests. Jensen (1976) argued that test bias should be studied against two criteria: predictive validity and construct validity. Reynolds and Ramsay (2003) differentiated a true group difference and test bias:

> In statistics, *bias* refers to systematic error in the estimation of a value. A biased test is one that systematically overestimates or underestimates the value of the variable it is intended to assess. If this bias occurs as a function of a nominal cultural variable, such as ethnicity or gender, cultural test

> bias is said to be present. On the Wechsler series of intelligence tests, for example, the difference in mean scores for Black and White American hovers around 15 points. If this figure represents a true difference between the two groups, the tests are not biased. If, however, the difference is due to systematic underestimation of the intelligence of Black Americans or overestimation of the intelligence of White Americans, the tests are said to be culturally biased (p. 68).

The line of research on test bias is much broader than the present study in that a key point in test bias research is to present argumentation, in addition to observed group differences, that bias exists (or does not exist). In this study, we did not attend to the explanations of possible group differences at the test level. Instead, we focus on group differences themselves resulted from different scoring methods. Furthermore, we did not test individual item bias or DIF; rather, we tested different sets of item parameters (e.g., testing for all item loadings/discriminations simultaneously), in accordance with measurement invariance research in the factor analysis framework. However, we acknowledge that testing for item bias is possible with the data used in this study. Wu and colleagues (2012) conducted analysis to examine DIF on a subscale of the Teacher Observation of Classroom Adaptation-Revised (TOCA-R), an earlier version of the scale used in the present study.

When factor analysis and IRT models are used to examine group mean differences in the latent factor scores or ability/proficiency scores, it is important to evaluate measurement invariance across groups. Measurement invariance of psychological constructs is a fundamental consideration in psychometrics. As early as 1968, Lord and Novick discussed invariant item parameters in their dealing with the relations between examinees' mental traits and test performance for different groups. There are also more recent scholarly articles on measurement invariance (e.g., Dimitrov, 2010; Kim & Yoon, 2011; Wu, Li, & Zumbo, 2007). In the CFA framework, if a particular factor model in each group well fits the data, it is important to test different levels of invariance. With continuous indicators, Gregorich (2006) described the implications when different types of invariance exist. If the interest is group differences in latent factor scores, indicator intercept parameters and factor loading parameters should be invariant across groups. Furthermore, the equality of factor variances (and factor covariances if more than one latent factor) can be tested and group mean differences examined accordingly. In the CFA framework, with ordered categorical indicators, indicator threshold parameters and factor loading parameters should be invariant across groups to compare group differences in latent factor scores (Muthén & Muthén, 2009; see Millsap & Yun-Tein, 2004, for a technical description of measurement invariance involving ordered categorical indicators). Similarly, factor variances and covariances can be compared between groups before factor means are compared.

For IRT models, measurement invariance is usually examined via the differential functioning of items and tests (DFIT) approach (Millsap, 2011, p. 223). Since IRT models are usually designed for item-level analysis, DIF is commonly assessed. An item with significantly different estimates for a parameter in multiple groups is considered a DIF item. DIF in IRT focuses on individual items. In this study, we focused on test-level differences in order to be consistent with the other scoring methods and to control for Type I error (see Banks, 2013, for more discussion on the benefits of analyzing bundles of items together instead of individual items for DIF). Measurement invariance in CFA, while could be used to test invariance for different items, is usually examined at increasingly stringent levels for sets of parameters for all items together (however, it is possible to have partial measurement invariance where not all parameters in the same set are invariant. See

Byrne, 2010, pp. 198–199, for a description of a testing strategy of invariance). Commonly used sets of parameters include factor loadings, factor intercepts/thresholds, and factor variances and covariances. For a technical guide on measurement invariance in both factor analysis and IRT frameworks, see Millsap (2011). Table 1 lists the comparisons among the five scoring methods.

In this study, we used scores based on the five methods mentioned earlier to examine group differences in the three subscales of the Teacher Observation of Classroom Adaptation Checklist (TOCA-C; Koth, Bradshaw, & Leaf, 2009). TOCA-C and its earlier versions are widely used in intervention studies to measure children's behaviors (e.g., Werthamer-Larsson, Kellam, & Wheeler, 1991). We were interested in mean differences between boys and girls, and between those with free/reduced-price lunch and those without. The magnitude of group differences was calculated using Cohen's *d* (Cohen, 1988). For the fifth scoring method only, group differences on the group means were divided by a scaling factor $D = 1.7$ to put the logits on the probit scale.

METHOD

Participants

Thirty-four teachers from three elementary schools in a large Midwestern school district participated in the study by completing the TOCA-C in reference to 577 students from Kindergarten to third grade in October 2010. The majority of students in two schools were African American (98%), while students in the third school were mostly Caucasian (59%; African American 39%). Of those 577 students in this study, 49.6% were girls and 37.8% received free/reduced-price lunch at school.

Measure

In this study, the 21-item TOCA-C (Koth, et al., 2009) was used to measure students' classroom behaviors. There are three subscales in the TOCA-C: seven items for the concentration problems subscale, nine items for the disruptive behavior subscale, and five items for the prosocial behavior subscale. Each student was rated on a 6-point scale (from *never* to *almost always*) for each behavior on the checklist. When examining internal consistency, the alpha coefficients for all TOCA-C subscales varied between .87 and .97 (Koth et al., 2009). The TOCA-C is a checklist version of the TOCA-R (Werthamer-Larsson, et al., 1990), which uses a structured interview format. Previous research demonstrates few quantitative differences when using either the checklist or structured-interview formats (Koth et al., 2009). Furthermore, TOCA-C scores indicated boys were typically rated as having more severe problem behaviors when compared with girls, which is consistent with previous findings (Bradshaw, Schaeffer, Petras, & Ialongo, 2010; Petras et al., 2005; Wu et al., 2012). Although all 21 items are created on a 6-point scale, some response options are rarely selected (e.g., "almost always" harms others). Because low frequencies may be a threat to the bivariate normality assumption of polychoric correlations (Muthén, 1993), some response options were collapsed. The scoring method recommended by Koth and colleagues (2009) is the average score method without response category collapsing.

TABLE 1
Comparisons of Five Scoring Methods

	Sum score method	*CFA with continuous indicators*	*CFA with ordered categorical indicators using mean- and variance-adjusted weighted least squares estimator*	*CFA with ordered categorical indicators using full information maximum likelihood-probit link*	*Multidimensional graded response item response theory model–logit link*
Data modeled	Sum scores; standard deviations of sum scores	Means, variances and covariances of items	Tetrachoric or polychoric correlations	Response patterns	Response patterns
Assumptions for scoring	True-score equivalent items	Equidistance between item response categories; items are normally distributed (nonnormality can be corrected with robust standard errors and adjusted chi-square statistics)	Bivariate normality of underlying latent response variables for tetrachoric and polychoric correlations	Consistency between response patterns and model specifications	Consistency between response patterns and model specifications
Assumptions for group mean comparisons	Sum scores normally distributed; equal variance across groups	Different levels of measurement invariance: factor loadings, item intercepts, item residuals, factor variances, factor covariances	Different levels of measurement invariance: factor loadings; thresholds; item residuals (with the Theta parameterization in Mplus), factor variances; factor covariances	Items function in a similar way across groups	Items function in a similar way across groups

Procedure

Data used in this study come from a large-scale group randomized trial evaluating the intervention efficacy of the Incredible Years Teacher Classroom Management Training (Webster-Stratton, 1994). General education classroom teachers from kindergarten to third grade in three schools as well as their classroom students were recruited in the trial. Every invited teacher consented to participate. Parent consent was obtained for 83% of these students, and each student with parental consent also agreed to participate in this study. At the beginning of the school year, participant teachers completed the TOCA-C in reference to each participating student's behaviors in order to measure the level of students' maladaptive and adaptive behaviors in the classroom from teachers' perspectives. At that time of TOCA-C data collection, the Incredible Years Teacher Classroom Management Training was not implemented yet. All study procedures were approved by the University of Missouri institutional review board.

Statistical Analysis

In this study, students in each classroom were rated by the same teacher and classrooms were sampled from three different schools. Therefore, the data used in this study have a complex structure and are not independent. This nonindependence was accounted for by specifying the schools as different strata and classrooms as clusters in Mplus for the second (using MLR estimator) and third (using WLSMV estimator) scoring methods. Because multiple group analysis is not available with the FIML estimator for ordered categorical variables in Mplus (fourth scoring method), mixture modeling with KNOWNCLASS was used (i.e., a latent categorical variable with two classes was specified but the class membership was known) as in Bovaird and Koziol (2012). This usage results in several model specifications that are different from the WLSMV estimator: (a) complex data structure is not considered (i.e., no stratum or cluster variable was used); (b) equality of item residuals across groups cannot be tested; and (c) there is one additional parameter for the latent categorical variable, estimated as the mean for the first latent class. In addition, since this method does not use a correlation matrix but response patterns directly, global model fit statistics are not available. Because of the large frequency table, the chi-square statistics testing the discrepancy between observed data and the model that would have been available if the model was simpler were not calculated for the dataset in the present study. For the fifth scoring method, it should be noted that the nonindependence did not exist at the item level (i.e., items are not clustered as in testlet models), but rather at the person level (i.e., persons are clustered because of the same teacher rating students in the same class). Therefore, the *confirmatory* item factor analysis model was used where each item measured only one dimension and the multidimensional graded response model was used. Similar to the fourth scoring method, the complex data structure was not considered because of the limitation of IRTPRO. Regular global fit statistics such as CFI and RMSEA are not available; instead, the –2*loglikelihood (–2LL) and the BIC statistics are reported. For the same model using the third and fifth scoring methods, the number of parameters to estimate is the same. For the same model using the fourth and fifth scoring methods, the former method had one more parameter to estimate.

RESULTS

Descriptive Statistics

Table 2 presents the frequency of each behavior and collapsing of categories for some items. The original response options were coded on a 1–6 scale. The specific coding depends on the item and the subscale so that a higher value for a concentration problems or disruptive behavior item indicates more problematic behavior and a higher value for a prosocial behavior item indicates more prosocial tendency. For example, the item "Works hard" measures concentration problems and is coded as 1 = *almost alway*s, and 6 = *never*; the item "Is friendly" measures prosocial behavior and is coded as 1 = *never* and 6 = *almost always*. The collapsing of categories was based on examining the frequency distribution so that after collapsing, there was no category with a very small size. After collapsing of categories, the new categories were coded on a scale starting from 0, instead of 1 to be consistent with what Mplus uses. For example, for the item "Concentrates," the two categories "never" and "rarely" were collapsed, and the new coding was 0 = *almost always*, 1 = *very often*, 2 = *often*, 3 = *sometimes*, and 4 = *rarely or never*. If no categories were collapsed for an item, one was subtracted from the original coding so that the item was on a 0–5 scale. The descriptive statistics before and after collapsing categories are in Table 3.

Measurement Invariance

We tested measurement invariance across groups for the second, third, fourth, and fifth scoring methods. The second scoring method assumes continuous item scores, following Gregorich (2006), configural invariance, metric invariance, scalar invariance, item residual invariance and factor variance and covariance invariance were tested. The third, fourth, and fifth scoring methods assume ordered categorical item scores. For the third scoring method, configural invariance, scalar invariance, item residual invariance, and factor variance and covariance invariance were tested. For the fourth and fifth scoring methods, configural invariance, scalar invariance, and factor variance and covariance invariance were tested. Scalar invariance in all four methods is a prerequisite before latent factor means can be compared. Model fit statistics for the four scoring methods are in Tables 4 and 5, for gender group comparisons and for free/reduced-price lunch status group comparisons, respectively.

Gender Groups

On the basis of results from the second scoring method (CFA with continuous indicators and MLR estimator), the configural invariance model only fit the data marginally acceptably (CFI = .91, RMSEA = .072 with 90% CI [.066, .077]) (the 90% CI for RMSEA is the default report in Mplus). The model fit statistics were slightly worse than those obtained by Koth and colleagues (2009; in Koth et al., there were two independent random samples for spring data and two independent random samples for fall data, resulting in four samples. CFI for the four samples ranged from .91 to .93; and RMSEA ranged from .06 to .07. However, data for boys and girls were analyzed together in Koth et al.). Metric invariance assuming equal factor loadings across gender groups, did not fit the data well (CFI = .90, RMSEA = .075 with 90% CI [.069,

TABLE 2
Frequency of Item Responses and Collapsing of Categories

	Almost always	*Very often*	*Often*	*Sometimes*	*Rarely*	*Never*	*Collapsing of categories*
CP1. Concentrates	129	149	130	112	51	6	(never, rarely)
CP2. Pays attention	132	152	122	106	57	8	(never, rarely)
CP3. Works hard	173	160	99	107	33	4	(never, rarely)
CP4. Stays on task	121	160	122	104	61	9	(never, rarely)
CP5. Is easily distracted	36	55	59	166	184	76	—
CP6. Completes assignments	224	179	79	63	28	4	(never, rarely)
CP7. Learns up to ability	173	159	95	103	39	6	(never, rarely)
DB1. Breaks rules	9	27	30	147	203	160	(almost always, very often)
DB2. Doesn't get along with others	16	18	24	98	174	247	—
DB3. Harms others	0	2	6	33	52	483	(very often, often, sometimes)
DB4. Gets angry when provoked by other children	10	26	35	104	183	218	—
DB5. Yells at others	2	6	26	72	143	328	(almost always, very often, often)
DB6. Fights	1	2	5	32	73	462	(almost always, very often, often, sometimes)
DB7. Lies	4	6	19	74	135	337	(almost always, very often, often)
DB8. Harms property	7	3	5	39	54	469	(almost always, very often, often, sometimes)
DB9. Teases classmates	2	10	18	93	139	312	(almost always, very often)
PB1. Is friendly	234	158	121	56	7	1	(never, rarely, sometimes)
PB2. Is liked by classmates	187	194	112	72	11	1	(never, rarely)
PB3. Shows empathy and compassion for others' feeling	123	139	114	134	56	9	(never, rarely)
PB4. Is rejected by classmates	6	3	8	81	177	302	(almost always, very often, often, sometimes)
PB5. Has many friends	163	154	128	92	27	12	—

Note. CP = concentration problems; DB = disruptive behavior, PB = prosocial behavior.

TABLE 3
Descriptive Statistics Before and After Collapsing Categories

	Before collapsing categories			*After collapsing categories*		
	M	*Skewness*	*Kurtosis*	M	*Skewness*	*Kurtosis*
CP1	2.70	0.32	−0.86	1.69	0.25	−1.04
CP2	2.70	0.37	−0.85	1.69	0.28	−1.07
CP3	2.44	0.51	−0.79	1.44	0.45	−0.98
CP4	2.74	0.38	−0.83	1.73	0.28	−1.06
CP5	2.90	0.68	−0.25	1.90	0.68	−0.25
CP6	2.14	0.96	0.08	1.13	0.89	−0.21
CP7	2.47	0.55	−0.75	1.46	0.47	−0.98
DB1	2.28	0.98	0.85	1.27	0.78	0.14
DB2	2.03	1.41	1.81	1.03	1.41	1.81
DB3	1.25	2.83	8.30	0.23	2.33	4.12
DB4	2.13	1.15	0.92	1.13	1.15	0.92
DB5	1.69	1.47	1.93	0.67	1.16	0.31
DB6	1.29	2.84	9.78	0.27	2.06	3.02
DB7	1.67	1.66	2.94	0.64	1.19	0.38
DB8	1.34	3.24	12.08	0.28	2.03	2.63
DB9	1.75	1.37	1.68	0.74	1.27	1.13
PB1	4.96	−0.72	−0.34	1.97	−0.56	−0.92
PB2	4.82	−0.65	−0.39	2.82	−0.62	−0.52
PB3	4.19	−0.23	−0.97	2.21	−0.13	−1.18
PB4	5.30	−1.70	4.18	1.35	−0.69	−0.93
PB5	4.52	−0.60	−0.36	3.52	−0.60	−0.36
	Cronbach's alpha before collapsing			Cronbach's alpha after collapsing		
CP	0.955			0.956		
DB	0.900			0.902		
PB	0.898			0.898		

Note. Item scores are on a 1–6 scale before collapsing categories and are on a 0–5 scale after collapsing. CP = concentration problems; DB = disruptive behavior, PB = prosocial behavior.

.080]) and modification indexes also suggested releasing equal loading constraints and cross-loading items (reflecting the necessity to modify the configural invariance model). However, if the configural invariance model were assumed correctly specified, then further constraining equality across gender groups in factor loadings, item intercepts, item residuals, and factor variances and covariances did not result in significantly worse model fit (chi-square tests for comparing nested models are in Table 4, the smallest p value was .338). That is, of the invariance models tested, Model 5 in Table 4 was the best in terms of model fit and parsimony (however, none of the models had very good model fit statistics; for Model 5, CFI = .89, RMSEA = .075 with 90% CI [.070, .081]).

On the basis of results from the third scoring method (CFA with ordered categorical indicators and WLSMV estimator), the fit of all tested models was good (CFI for all four models were .99; RMSEA ranged from .048 to .062). The scalar invariance model did not fit significantly worse than the configural model (Model 1 vs. Model 3, $\Delta\chi^2(71) = 83.5$, $p = .147$). However, further constraining item residuals resulted in significantly worse model fit (Model 3 vs. Model 4,

TABLE 4
Model Fit Statistics for Evaluating Measurement Invariance Across Gender Groups

	CFA with continuous indicators (MLR [a])				*CFA with ordered categorical indicators using WLSMV*				*CFA with ordered categorical indicators using FIML; probit link*			*Multidimensional graded response IRT model using BAEM; logit link*			
	# Para.	χ^2(df) [b]	*CFI*	*RMSEA (90% CI)*[c]	*# Para.*	χ^2(df) [b]	*CFI*	*RMSEA (90% CI)* [c]	*# Para.*	*LL*	*Sample adjusted BIC*	*# Para.*[d]	*−2LL*	*AIC*	*BIC*
M1	132	922.0 (372)	.91	.072 (.066, .077)	202	780.2 (372)	.99	.062 (.056, .068)	203	−11127.1	22900.4	202	21469.6	21873.6	22753.9
M2	111	1024.0 (393)	.90	.075 (.069, .080)											
M3	93	1084.4 (411)	.89	.075 (.070, .081)	131	810.3 (443)	.99	.054 (.048, .059)							
M4	72	1117.0 (432)	.89	.074 (.069, .079)	110	795.7 (464)	.99	.050 (.044, .056)	111	−11207.4	22768.1	110	21623.3	21843.3	22322.6
M5	**66**	**1156.7 (438)**	**.89**	**.075 (.070, .081)**	**125**	**747.6 (449)**	**.99**	**.048 (.042, .054)**	**105**	**−11210.3**	**22754.8**	**104**	**21644.3**	**21852.3**	**22305.6**
		$\Delta\chi^2(\Delta df)$	*p*			$\Delta\chi^2(\Delta df)$ [e]	*p*			Δ–2LL(Δ*df*)	*p*			Δ–2LL(Δ*df*)	*p*
M1 vs. M2		4.7 (21)	1.000		M1 vs. M3	83.5 (71)	.147		M1 vs. M4	160.6 (92)	<.001	M1 vs. M4		153.7 (92)	<.001
M2 vs. M3		3.5 (18)	1.000		M3 vs. M4	48.9 (21)	<.001		M4 vs. M5	5.8 (6)	.447	M4 vs. M5		21.1 (6)	.002
M3 vs. M4		2.0 (21)	1.000		M3 vs. M5	7.0 (6)	.317		M1 vs. M5	166.4 (98)	<.001	M1 vs. M5		174.8 (98)	<.001
M4 vs. M5		6.8 (6)	.338												

Note. MLR = maximum likelihood estimator with robust standard errors and adjusted chi-square statistics; WLSMV = mean- and variance-adjusted weighted least squares estimator; FIML = full information maximum likelihood estimator; IRT = item response theory; BAEM = Bock-Aitkin expectation-maximization estimation procedure; LL = loglikelihood. M1, no constraints (configural invariance model); M2, factor loadings constrained (metric invariance model); M3, factor loadings and intercepts/thresholds constrained (scalar invariance model); M4, item residuals constrained; M5, factor variances and covariances constrained (for CFA with continuous indicators, item residuals were constrained as in M4; for the WLSMV method, item residuals were fixed at ones for the male group for identification purpose and freed in the female group; for the FIML and IRT methods, item residuals were not parameterized). Models in bold are used to examine gender group differences.

[a]Yuan-Bentler adjusted test statistic.

[b]$p < .001$.

[c]Mplus reports the 90% confidence interval for RMSEA by default.

[d]One more parameter than the WLSMV method because of estimation of a latent categorical variable mean.

[e]Calculated based on DIFFTEST command in Mplus.

TABLE 5
Model Fit Statistics for Evaluating Measurement Invariance Across Free or Reduced-Price Lunch Status Groups

	CFA with continuous indicators (MLR [a])				*CFA with ordered categorical indicators using WLSMV*				*CFA with ordered categorical indicators using FIML; probit link*			*Multidimensional graded response IRT model using BAEM; logit link*			
	# Para.	$\chi^2(df)$ [b]	*CFI*	*RMSEA (90% CI)* [c]	*# Para.*	$\chi^2(df)$ [b]	*CFI*	*RMSEA (90% CI)* [c]	*# Para.*	*LL*	*Sample adjusted BIC*	*# Para.* [d]	*−2LL*	*AIC*	*BIC*
M1	132	923.6 (372)	.92	.072 (.066, .078)	202	737.5 (372)	.99	.058 (.052, .065)	203	−11138.5	22923.2	202	21518.7	21922.7	22803.0
M2	111	989.4 (393)	.91	.073 (.067, .078)											
M3	93	1017.1 (411)	.91	.071 (.066, .077)	131	762.0 (443)	.99	.050 (.044, .056)							
M4	72	1053.5 (432)	.91	.071 (.065, .076)	**110**	**720.5 (464)**	**.99**	**.044 (.037, .050)**	111	−11208.8	22771.0	110	21667.1	21887.1	22366.4
M5	**66**	**1072.7 (438)**	**.90**	**.071 (.066, .076)**	125	762.1 (449)	.99	.049 (.043, .055)	**105**	**−11216.4**	**22767.1**	**104**	**21682.2**	**21890.2**	**22343.4**
		$\Delta\chi^2(\Delta df)$	*p*			$\Delta\chi^2(\Delta df)$ [e]	*p*			$\Delta -2LL(\Delta df)$	*p*			$\Delta -2LL(\Delta df)$	*p*
M1 vs. M2		3.17 (21)	1.00	M1 vs. M3		66.6 (71)	.625	M1 vs. M4		140.6 (92)	<.001	M1 vs. M4		148.4 (92)	<.001
M2 vs. M3		1.40 (18)	1.00	M3 vs. M4		29.2 (21)	.109	M4 vs. M5		15.2 (6)	.018	M4 vs. M5		15.1 (6)	.019
M3 vs. M4		1.96 (21)	1.00	M3 vs. M5		22.0 (6)	.001	M1 vs. M5		155.8 (98)	<.001	M1 vs. M5		163.5 (98)	<.001
M4 vs. M5		3.46 (6)	.749												

Note. MLR = maximum likelihood estimator with robust standard errors and adjusted chi-square statistics, WLSMV = mean- and variance-adjusted weighted least squares estimator, FIML = full information maximum likelihood estimator, IRT = item response theory, BAEM = Bock-Aitkin expectation-maximization estimation procedure, LL = loglikelihood. M1, no constraints (configural invariance model); M2, factor loadings constrained (metric invariance model); M3, factor loadings and intercepts/thresholds constrained (scalar invariance model); M4, item residuals constrained; M5, factor variances and covariances constrained (for CFA with continuous indicators, item residuals were constrained as in M4; for the WLSMV method, item residuals were fixed at ones for the male group for identification purpose and freed in the female group; for the FIML and IRT methods, item residuals were not parameterized). Models in bold are used to examine differences between the free or reduced-price lunch group and the group without free or reduced-price lunch.

[a]Yuan-Bentler adjusted test statistic.

[b]$p < .001$.

[c]Mplus reports the 90% confidence interval for RMSEA by default.

[d]One more parameter than the WLSMV method because of estimation of a latent categorical variable mean.

[e]Calculated based on DIFFTEST command in Mplus.

$\Delta\chi^2(21) = 48.9, p < .001$). If factor variance and covariance constraints were added to the scalar invariance model, the model fit did not decrease significantly (Model 3 vs. Model 5, $\Delta\chi^2(6) = 7.0, p = .317$). Therefore, the model (Model 5) with invariances of factor loading, item thresholds, and factor variance and covariance was the best in terms of model fit and parsimony.

Using the fourth scoring method (CFA with ordered categorical indicators and FIML estimator), the scalar invariance model fit significantly worse than the configural invariance model (Model 1 vs. Model 4, Δ–2LL = 160.6, $df = 92, p < .001$), suggesting that items do not function in the same way across the two gender groups after controlling for true scores (i.e., there were DIF items). Since no latent response variables were involved in this method, residual variances cannot be modeled like in the WLSMV method. However, further constraining factor variance and covariance did not decrease model fit significantly (Model 4 vs. Model 5, Δ–2LL = 5.8, $df = 6, p = .447$). Comparing results from the WLSMV method and the FIML, the significant difference between the configural invariance model and the scalar invariance model using FIML was possibly a result from different parameterization. Furthermore, on the basis of model fit indexes that take model complexity into consideration (e.g., AIC, BIC, and sample-size adjusted BIC), the model with equal factor loading and item threshold constraints, and equal factor variance and covariance constraints (i.e., Model 5) was the best.

Results from the fifth scoring method (graded response model with BAEM estimation method) suggested that the scalar invariance model fit significantly worse than the configural invariance model (Model 1 vs. Model 4, Δ–2LL = 153.7, $df = 92$, $p < .001$), similar to the results in the fourth method. Further constraining factor variance and covariance decreased model fit significantly at the .05 level but not at the .001 level (Model 4 vs. Model 5, Δ–2LL = 21.1, $df = 6, p = .002$). The scalar invariance model (Model 4) had better fit based on AIC and the variance and covariance invariance model (Model 5) had better fit based on BIC. Later group mean comparisons were based on results from Model 5.

The fourth and fifth scoring methods were the same IRT model (multidimensional graded response model) with different estimation algorithms, link functions, and parameterizations (differences between the two program, Mplus for the fourth method and IRTPRO for the fifth method are outside the discussion here). Besides the difference of an additional parameter in the fourth scoring method, item thresholds were parameters in the fourth scoring method while item intercepts were parameters in the fifth scoring method.

Free/Reduced-Price Lunch Status Groups

Measurement invariance results from the second scoring method (CFA with continuous indicators and MLR estimator) were similar to those for the gender groups (CFI = .92, RMSEA = .072 with 90% CI [.066, .078], see Table 5). That is, the configural invariance model only fit the data marginally acceptably. However, if the configural invariance model were assumed correctly specified, then further constraining equality across lunch status groups in factor loadings, item intercepts, item residuals, and factor variances and covariances did not result in significant worse model fit (chi-square tests for comparing nested models are in Table 5, the smallest p value was .749). Of the invariance models tested, Model 5 in Table 5 was the best in terms of model fit and parsimony (however, none of the models had very good model fit statistics; for Model 5, CFI = .90, RMSEA = .071 with 90% CI [.066, .076]).

Based on results from the third scoring method (CFA with ordered categorical indicators and WLSMV estimator), the fit of all tested models was good (CFI for all four models were .99; RMSEA ranged from .044 to .058). The scalar invariance model did not fit significantly worse than the configural model (Model 1 vs. Model 3, $\Delta\chi^2(71) = 66.6$, $p = .625$, for chi-square difference test). Further constraining item residuals did not decrease model fit significantly (Model 3 vs. Model 4, $\Delta\chi^2(21) = 29.2$, $p = .109$). If factor variance and covariance constraints were added to the scalar invariance model, the model fit decreased significantly (Model 3 vs. Model 5, $\Delta\chi^2(6) = 22.0$, $p = .001$). Therefore, the model with invariances of factor loading, item thresholds, and item residuals (Model 4) was the best in terms of model fit and parsimony.

Using the fourth scoring method (CFA with ordered categorical indicators and FIML estimator), measurement invariance results for the lunch status groups were similar to those for the gender groups. The scalar invariance model fit significantly worse than the configural invariance model (Model 1 vs. Model 4, Δ–2LL = 140.6, $df = 92$, $p < .001$), suggesting that items do not function in the same way across the two gender groups after controlling for true scores (i.e., there were DIF items). Because no latent response variables were involved in this method, residual variances cannot be modeled like in the WLSMV method. However, further constraining factor variance and covariance did not decrease model fit significantly (Model 4 vs. Model 5, Δ–2LL = 15.2, $df = 6$, $p = .018$). On the basis of sample-size adjusted BIC, the model with equal factor loading and item threshold constraints, and equal factor variance and covariance constraints (i.e., Model 5) was the best.

Using the fifth scoring method (graded response model with BAEM estimation method), measurement invariance results for the lunch status groups were similar to those for the gender groups. The scalar invariance model fit significantly worse than the configural invariance model (Model 1 vs. Model 4, Δ–2LL = 148.4, $df = 92$, $p < .001$), similar to the results in the fourth method. Further constraining factor variance and covariance did not decrease model fit significantly (Model 4 vs. Model 5, Δ–2LL = 15.1, $df = 6$, $p = .019$). The scalar invariance model (Model 4) had better fit based on AIC and the variance and covariance invariance model (Model 5) had better fit based on BIC. Later group mean comparisons were based on results from Model 5.

Group Mean Differences

Gender Groups

Gender differences were examined with the five scoring methods. For the first scoring method, gender differences were examined using observed sum scores for the three subscales of TOCA-C. For the second and third scoring methods, the best fitting model based on measurement invariance results were used. For the fourth and fifth scoring methods, the factor variance and covariance model (Model 5) was used as an illustration. With the fourth scoring method, the configural invariance model was not useful in comparing gender differences because all factor means were fixed at zero for model identification purpose. In the final models of the second, third, fourth and fifth scoring methods, factor variances were fixed at one and the factor mean was fixed at zero for the male group. Therefore, the factor mean for the female group represents the gender difference, as well as the effect size (in terms of Cohen's *d*) of gender differences. The effect size

of gender differences based on the first scoring method was calculated using the pooled within-group standard deviation of sum scores for each subscale. Table 6 displays those differences. A negative mean difference indicates a lower mean for girls than for boys.

All five scoring methods suggested gender differences in all three TOCA-C subscales ($ps < .001$, see Table 6). Those gender differences were consistent with previous findings that boys had more problem behavior and less prosocial behavior than girls (Greener & Crick, 1999; Ostrov, Crick, & Keating, 2005; Zimmer-Gembeck, Geiger, & Crick, 2005). The effect sizes of gender difference in each of the subscales using the five scoring methods differed but not much. The largest difference between scoring methods was 0.18, or 25%, and was between the second and fifth scoring methods for concentration problems. We reported these gender differences using Cohen's (1988) suggestions for effect sizes. For concentration problems, gender differences were medium to large (Cohen's d ranged from –0.72 to –0.54); the second scoring method (using MLR estimator) demonstrated the largest gender difference, followed by the third method (using WLSMV estimator). For disruptive behavior, gender differences were small to medium (Cohen's d ranged from –0.46 to –0.37); the fourth (FIML) method demonstrated the largest gender difference, followed by the third (WLSMV) method. For prosocial behavior, gender differences were small to medium (Cohen's d ranged from 0.35 to 0.40); the third (WLSMV) and fourth (FIML) methods demonstrated the largest gender differences, followed by the sum score method. The first two methods assume that item scores are on an interval scale. However, no consistent pattern was observed regarding which of the two would result in larger effect sizes of gender difference. The last three methods consider the ordered categorical nature of the item responses; but again, no consistent pattern was observed regarding which one is associated with largest effect sizes.

Free/Reduced-Price Lunch Status Groups

Table 7 displays mean differences on the three TOCA-C subscales based on different scoring methods. For the first scoring method, group differences were examined using observed sum scores for the three subscales. For the second, third, fourth, and fifth scoring methods, group differences were calculated based on their respective models in bold as reported in Table 5. Mean differences bases on the second, fourth and fifth scoring methods were also their effect sizes. Mean differences based on the third scoring method were standardized in the Cohen's d metric using information of pooled within group standard deviations of the latent factors. A negative mean difference indicates a lower mean for the free/reduced-price lunch group than for the group without free/reduced-price lunch.

The statistical significance levels of these mean differences differed as well as their magnitudes. The largest difference between scoring methods was 0.24, or 67%, and was between the first and fifth scoring methods for disruptive behavior. Using Cohen's (1988) suggestions for effect sizes, on the basis of results from the first four scoring methods, the differences between the free/reduced-price lunch group and the group without free/reduced-price lunch were small, or small to medium in the three TOCA-C subscales (Cohen's d ranged from 0.17 to 0.23 for concentration problems, 0.29 to 0.36 for disruptive behavior, and –0.26 to –0.23 for prosocial behavior. However, not all differences were statistically significant with the four methods). In terms of statistical significance levels, the sum score method (i.e., the first method) seemed to have the largest power to detect group differences, as indicated by the smallest p value for each of the three subscales. However,

TABLE 6
Gender Differences

	Sum score method			*CFA with continuous indicators (MLR)*		*Graded response model (WLSMV)*		*Graded response model (FIML)*		*Multidimensional graded response IRT model using BAEM; logit link*	
	Mean difference	*Cohen's* d	p	*Mean difference*[a]	p	*Mean difference*[a]	p	*Mean difference*[a]	p	*Mean difference*[ab]	p
Concentration problems	−4.92	−0.64	<.001	−0.72	<.001	−0.67	<.001	−0.64	<.001	−0.54	<.001
Disruptive behavior	−2.30	−0.37	<.001	−0.37	<.001	−0.39	.008	−0.46	<.001	−0.38	<.001
Prosocial behavior	1.80	0.39	<.001	0.35	<.001	0.40	.004	0.40	<.001	0.35	.003

Note. A negative mean difference indicates a lower mean for girls than for boys. MLR = maximum likelihood estimator with robust standard errors and adjusted chi-square statistics, WLSMV = mean- and variance-adjusted weighted least squares estimator, FIML = full information maximum likelihood estimator, IRT = item response theory, BAEM = Bock-Aitkin expectation-maximization estimation procedure.

[a]Also effect size as Cohen's *d*.

[b]The ability/proficiency scores are estimated with a logistic ogive. Group mean differences are divided by a scaling factor of 1.7.

TABLE 7
Free/Reduced-Price Lunch Status Differences

	Sum score method			*CFA with continuous indicators (MLR)*		*Graded response model (WLSMV)*			*CFA with ordered categorical indicators using FIML*		*Multidimensional graded response IRT model using BAEM; logit link*	
	Mean difference	*Cohen's* d	p	*Mean difference*[a]	p	*Mean difference*	*Cohen's* d	p	*Mean difference*[a]	p	*Mean difference*[ab]	p
Concentration problems	1.64	0.20	.018	0.23	.036	0.19	0.18	.116	0.17	.055	0.01	.505
Disruptive behavior	2.24	0.36	<.001	0.33	.003	0.31	0.29	.067	0.30	.001	0.12	.549
Prosocial behavior	−1.13	−0.24	.005	−0.26	.039	−0.26	−0.24	.152	−0.23	.010	−0.08	.470

Note. A negative mean difference indicates a lower mean for the free/reduced-price lunch group than for the group without free/reduced-price lunch. MLR = maximum likelihood estimator with robust standard errors and adjusted chi-square statistics; WLSMV = mean- and variance-adjusted weighted least squares estimator; FIML = full information maximum likelihood estimator; IRT = item response theory; BAEM = Bock-Aitkin expectation-maximization estimation procedure.

[a]Also effect size as Cohen's *d*.

[b]The ability/proficiency scores are estimated with a logistic ogive. Group mean differences are divided by a scaling factor of 1.7.

this should not be taken as evidence that this scoring method is the best. Results from the fifth method, the multidimensional graded response model using the BAEM estimator and the IRTPRO program, did not suggest any statistically significant group difference in the three subscales. It is interesting to note that although the fourth and fifth scoring methods are conceptually the same IRT model (with a difference in the link function which was corrected by the scaling factor $D = 1.7$) and the model fit statistics (e.g., loglikelihood, BIC) were similar (see Table 5; for example, the Δ–2LL statistic comparing M1 and M4 was 140.6 with the fourth scoring method and was 148.4 with the fifth scoring method), the conclusions about group mean differences were very different. This may be due to the differences in the estimators (FIML vs. BAEM) or the programs (Mplus vs. IRTPRO), or both.

DISCUSSION

In this study, we described five scoring methods in terms of their assumptions. We also examined group mean differences with an empirical example of data on TOCA-C by applying these scoring methods. The first scoring method (sum/average of item scores) is easy to implement. When constructs are well measured with sufficient internal consistency, this method seems appealing because of its simplicity. The other four methods consider the differences in contributions of item scores to measure a construct. With Likert-type items, considering the ordinality of responses and using the WLSMV estimator (i.e., the third scoring method) resulted in better model fit in this study than treating responses on a continuum and correcting for nonnormality (i.e., the second scoring method). This is consistent with results from earlier simulation studies (Beauducel & Herzberg, 2006). For the fourth scoring method, because of the lack of global model fit statistics and difficulty in evaluating the factor analysis model, factor scores of ordered categorical indicators using FIML does not seem appealing to applied researchers. However, because of the connection between CFA models with the FIML estimation method and IRT models, new methodological advancements are integrating the two frameworks and take advantages of both. As Edwards, Wirth, Houts, and Xi (2012) stated, "One advantage is that FIML can be applied directly to linear and nonlinear models without any ancillary steps (e.g., the tetrachoric-polychoric correlations) . . . Increasingly there are FIML options available for nonlinear models in popular SEM software" (p. 204). The fifth scoring method is within the IRT framework. Like the fourth method, this fifth method does not provide global model fit statistics that are usually available when summary statistics are used in the estimation process. Although not shown in the present study, advantages of this scoring method include more item-level diagnostics. Recent developments in multidimensional IRT models also allow different estimation methods that are more efficient than maximum likelihood estimators (e.g., Cai, 2010).

Results from this study suggest that factor analysis with ordered categorical indicators with the WLSMV estimator (i.e., the third scoring method) leads to better model fit. Earlier research based on a simulation study has also found that the WLSMV method outperforms the maximum likelihood method (Beauducel & Herzberg, 2006). However, ideally, researchers should not stop here but should continue to identify conditions when it does or does not apply. For example, Forero, Maydeu-Olivares, and Gallardo-Pujol (2009) have found that under some circumstances the unweighted least squares estimator produces more accurate results than the WLSMV (a different name, diagonally weighted least squares, or DWLS, is used in the cited article). Researchers are

also encouraged to apply those methods to examine psychometric properties of other scales that consist of Likert-type items.

We applied the five scoring methods to an empirical dataset and examined group mean differences in the results. It is possible to conduct simulation studies to systematically evaluate these methods. For example, Jöreskog and Moustaki (2001) compared three approaches for factor analysis with ordinal variables: (a) the underlying bivariate normal approach, (b) the normal ogive approach, and (c) the proportional odds model approach, using an empirical example and simulated data. The underlying bivariate normal approach in Jöreskog and Moustaki (2001) is conceptually similar to the WLS(MV) approach in this study (i.e., third scoring method) except that all parameters (thresholds for ordinal variables, factor loadings) are estimated simultaneously (this is not available in Mplus yet).

Our results show that gender differences were statistically significant and their effect sizes were close to medium in the three TOCA-C subscales using any of the scoring methods. Any method had sufficient power to detect such gender differences. However, for free/reduced-price lunch status group mean differences, the conclusions about statistical significances depended on the scoring method. The effect sizes of these group mean differences also varied by scoring method. From an evidence-collecting perspective, gender differences are easier to detect than free/reduced-price lunch status group differences in the TOCA-C subscales. It should be noted that these are concomitant of the TOCA-C, and do not necessarily generalize to another scale measuring a similar construct or scales measuring different constructs.

Researchers have been using multifactor CFA models for a long time. Many data and methodological considerations have been included in the development of those models including complex data structure, missing data, nonnormality, ordinal and count data, estimation methods, relations with traditional psychometrics statistics (e.g., reliability, validity, etc). Most data issues that applied researchers encounter have been addressed in the CFA framework. The availability of user friendly software has resulted in a lot of empirical studies. On the other hand, multidimensional IRT models have gained more attention recently because of development of new programs (e.g., Yao & Boughton, 2007). As Reckase (2009) stated, "It is difficult to separate the estimation procedures from the programs used to implement them. An excellent estimation methodology may perform poorly because it is improperly programmed or programmed in an inefficient way" (p. 137). Scoring methods could differ by estimators, algorithms used and computer programs. These differences likely result in different conclusions when it comes to the examination of individual and group differences. This study is a piece of evidence of such consequences.

Findings of the present study offer several implications and future research directions for educational researchers. First and foremost, scoring methods matter for substantive research. Although it is sometimes not practical to apply a scoring method that is consistent with the model used when a scale is developed, it is important for researchers to be aware of and to check the statistical assumptions—whether implicit or explicit—associated with the scoring method. For example, researchers have been well aware that it is important to report the reliability (e.g., Cronbach's alpha) of item scores when summing/averaging them to create a composite score. The behind-the-scene reason that involves true-score equivalent items, however, is less well known. Second, for Likert-type items, factor analysis with ordered categorical indicators with the WLSMV estimator may lead to better results in terms of model fit. The nature of this method puts it in advantage for ordered categorical variables. However, assessment of most model fit indexes is based on the more popular maximum likelihood estimation (e.g., Hu & Bentler, 1998, 1999). There

is a need to research on model fit indexes that are appropriate for different estimation methods. Third, besides the aforementioned need for simulation studies to systematically investigate which scoring method is better under certain conditions, researchers—both methodological and applied researchers—should work together to develop practical tools that are easy to use to implement sophisticated scoring methods. For example, new practical tools could be developed that easily provide individual scores consistent with item factor analysis models or multidimensional IRT models without complicated data management or programming.

AUTHOR NOTES

Ze Wang is an assistant professor of Educational Psychology with an emphasis on statistics and measurement in the Department of Educational, School and Counseling Psychology at the University of Missouri. Her research interests include statistical modeling and psychometrics. **David Rohrer** is a doctoral candidate in the Counseling Psychology program in the Department of Educational, School and Counseling Psychology at the University of Missouri. His research interests include youth depression and psychometrics. **Chi-ching Chuang** is a doctoral student in School Psychology in the Department of Educational, School and Counseling Psychology at the University of Missouri-Columbia. Her research interests include evaluation and intervention for children with behavioral and emotional difficulties. **Mayo Fujiki** is a doctoral candidate in the School Psychology program in the Department of Educational, School and Counseling Psychology at the University of Missouri. Her research interests include school-based program evaluation and consultation. **Keith Herman** is a professor of Counseling Psychology in the Department of Educational, School and Counseling Psychology at the University of Missouri. His research focuses on using a prevention science framework to promote children's mental health. **Wendy Reinke** is an associate professor of School Psychology in the Department of Educational, School and Counseling Psychology at the University of Missouri. Her research interests focus on developing and evaluating practices to promote effective classroom management.

FUNDING

The research reported here was partly supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305A100342 to the University of Missouri. The opinions expressed are those of the authors and do not represent views of the Institute of Education Sciences or of the U.S. Department of Education.

REFERENCES

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23. doi:10.1177/0146621697211001

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243–277). Mahwah, NJ: Erlbaum.

Banks, K. (2013). A synthesis of the peer-reviewed differential bundle functioning research. *Educational Measurement: Issues and Practice*, *32*, 43–55. doi:10.1111/emip.12004

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*, 186–203.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*, *46*, 443–459.

Bollen, K. A., Bauer, D. J., & Christ, S. L. (2010). Overview of structural equation models and recent extensions. In S. Kolenikov, D. Steinley, & L. Thombs (Eds.), *Statistics in the social sciences: Current methodological developments* (pp. 37–79). Hoboken, NJ: Wiley.

Bovaird, J. A., & Koziol, N. A. (2012). Measurement models for ordered-categorical indicators. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 495–511). New York, NY: Guilford Press.

Bradshaw, C. P., Schaeffer, C. M., Petras, H., & Ialongo, N. (2010). Predicting negative life outcomes from early aggressive–disruptive behavior trajectories: Gender differences in maladaptation across life domains. *Journal of Youth and Adolescence*, *39*, 953–966. doi:10.1007/s10964-009-9442-8

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (2nd ed.). New York, NY: Routledge.

Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*, 307–335. doi:10.3102/1076998609353115

Cai, L., du Toit, S. H. C., & Thissen, D. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]*. Chicago, IL: SSI International.

Cohen, J. (1988). *Statistical power analysis for the bahavioral sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Cooke, D. J., Kosson, D. S., & Michie, C. (2001). Psychopathy and ethnicity: Structural, item, and test generalizability of the Psychopathy Checklist—Revised (PCL-R) in Caucasian and African American participants. *Psychological Assessment*, *13*, 531–542. doi:10.1037/1040-3590.13.4.531

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, *43*, 121–149. doi:10.1177/0748175610373459

Dorman, J. P. (2003). Cross-national validation of the what is happening in this class? (WIHIC) questionnaire using confirmatory factor analysis. *Learning Environments Research*, *6*, 231–245. doi:10.1023/a:1027355123577

Edwards, M. C., Wirth, R. J., Houts, C. R., & Xi, N. (2012). Categorical data in the structural equation modeling framework. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 195–208). New York, NY: Guilford Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466–491. doi:10.1037/1082-989X.9.4.466

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, *16*, 625–641. doi:10.1080/10705510903203573

Greener, S., & Crick, N. R. (1999). Normative beliefs about prosocial behavior in middle childhood: What does it mean to be nice? *Social Development*, *8*, 349–363. doi:10.1111/1467-9507.00100

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, *44*, S78–S94.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijoff.

Harter, S. (1981). A new self-report scale of intrinsic versus extrinsic orientation in the classroom: Motivational and informational components. *Developmental Psychology*, *17*, 300–312. doi:10.1037/0012-1649.17.3.300

Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*, 424–453. doi:10.1037/1082-989x.3.4.424

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Jensen, A. R. (1976). Test bias and construct validity. *Phi Delta Kappan*, *58*, 340–346.

Jöreskog, K. G. (1990). New developments in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity*, *24*, 387–404. doi:10.1007/bf00152012

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, *36*, 347–387. doi:10.1207/s15327906347-387

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling*, *18*, 212–228. doi:10.1080/10705511.2011.557337

Koth, C. W., Bradshaw, C. P., & Leaf, P. J. (2009). Teacher Observation of Classroom Adaptation—Checklist: Development and factor structure. *Measurement and Evaluation in Counseling and Development*, *42*, 15–30. doi:10.1177/0748175609333560

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

McGrew, K. S., & Woodcock, R. W. (2001). Technical manual. *Woodcock-Johnson III*. Itasca, IL: Riverside.

McKeon, J. J. (1966). *Canonical analysis: Some relations between canonical correlation, factor analysis, discriminant function analysis, and scaling theory*. Sarasota, FL: Psychometric Society.

Millsap, R. E. (2011). *Statistical approach to measurement invariance*. New York, NY: Routledge.

Millsap, R. E., & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 380–392). New York, NY: Guilford Press.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*, 479–515.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, *17*, 131–154. doi:10.2307/1165166

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115–132.

Muthén, B. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–234). Newbury Park, CA: Sage.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*, 443–460. doi:10.1007/bf02296207

Osterlind, S. J., & Wang, Z. (2012). Item response theory in measurement, assessment and evaluation for higher education. In C. Secolsky & D. B. Denison (Eds.), *Handbook on measurement, assessment, and evaluation in higher education* (pp. 150–160). New York, NY: Routledge.

Ostrov, J. M., Crick, N. R., & Keating, C. F. (2005). Gender-biased perceptions of preschoolers' behavior: How much is aggression and prosocial behavior in the eye of the beholder? *Sex Roles*, *52*, 393–398. doi:10.1007/s11199-005-2681-6

Petras, H., Ialongo, N., Lambert, S. F., Barrueco, S., Schaeffer, C. M., Chilcoat, H., & Kellam, S. (2005). The utility of elementary school TOCA-R scores in identifying later criminal court violence among adolescent females. *Journal of the American Academy of Child & Adolescent Psychiatry*, *44*, 790–797. doi:10.1097/01.chi.0000166378.22651.63

Raykov, T. (2012). Scale construction and development using structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 472–492). New York, NY: Guilford Press.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Reynolds, C. R., & Ramsay, M. C. (2003). Bias in psychological assessment: An empirical review and recommendations. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology* (Vol. 10, pp. 67–93). Hoboken, NJ: Wiley.

Samejima, F. (1969). *Estimation of latent trait ability using a response pattern of graded scores* (Psychometrika Monograph No. 17). Richmond, VA: Psychometric Society.

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, *70*, 533–555.

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. (2009). Testing measurement invariance using multigroup CFA: differences between educational groups in human values measurement. *Quality & Quantity*, *43*, 599–616. doi:10.1007/s11135-007-9143-x

Webster-Stratton, C. (1994). *The Incredible Years Teacher Training Series*. Seattle, WA: Incredible Years.

Werthamer-Larsson, L., Kellam, S. & Oveson-McGregor, K. E. (1990). *Teacher interview: Teacher Observation of Classroom Adaptation-Revised (TOCA-R). Johns Hopkins Prevention Center Training Manual*. Baltimore, MD: Johns Hopkins University.

Werthamer-Larsson, L., Kellam, S., & Wheeler, L. (1991). Effect of first-grade classroom environment on shy behavior, aggressive behavior, and concentration problems. *American Journal of Community Psychology*, *19*, 585–602. doi:10.1007/bf00937993

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*, 58–79. doi:10.1037/1082-989x.12.1.58

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analyses: A demonstration with TIMSS data. *Practical Assessment, Research, and Evaluation*, *12*(3), 1–26.

Wu, J., King, K. M., Witkiewitz, K., Racz, S. J., & McMahon, R. J. (2012). Item analysis and differential item functioning of a brief conduct problem screen. *Psychological Assessment*, *24*(2), 444–454. doi:10.1037/a0025831

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, *31*, 83–105. doi:10.1177/0146621606291559

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological methodology*, *30*, 165–200. doi:10.1111/0081-1750.00078

Zimmer-Gembeck, M. J., Geiger, T. C., & Crick, N. R. (2005). Relational and physical aggression, prosocial behavior, and peer relations. *Journal of Early Adolescence*, *25*, 421–452. doi:10.1177/0272431605279841

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.